PB89-143762

Classification of Binary Mass Spectra of Toxic Compounds with an Inductive Expert System and Comparison with SIMCA Class Modeling (Journal Version)

(U.S.) Environmental Monitoring Systems Lab.
Research Triangle Park (NC)

1988

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before complet

1. REPORT NO. EPA/600/J-88/254	2.	3. PB89-143782
4. TITLE AND SUBTITLE Classification of Binary Mass Spectra of Toxic Compounds with an Inductive Expert System and Comparison with Simca Class Modeling (Journal Version)		5. REPORT DATE 1988 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Donald R. Scott		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS (Same as Box 12)		10. PROGRAM ELEMENT NO. DU-L104, Project A02/01 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS U.S. Environmental Protection Agency Environmental Monitoring Systems Laboratory Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA/600/09

15. SUPPLEMENTARY NOTES

Analytica Chimica Acta, 211:11-29, 1988

16. ABSTRACT

The performance of an inexpensive, inductive rule-building expert system, 1ST-CLASS, using the ID3 algorithm was compared to that of SIMCA class modeling in classifying the binary mass spectra of 78 toxic and related compounds. The compressed mass spectra consisted of 17 masses chosen using information theory. The expert rules verified the six main classes and two subclasses found with SIMCA class modeling. These classes were: all benzenes and all alkanes/alkenes (alkaenes); nonhalobenzenes, chlorobenzenes, bromoalkaenes, and chloroalkaenes; and mono-, di-chloroalkaenes and polychloroalkaenes. Training set classification accuracies obtained with the expert rules gave a classification accuracy of 97 to 100% vs. 79 to 96% for SIMCA. Predictive accuracy for the four main classes was 78%. In general fewer masses were involved with the rules than with the SIMCA models, and the rules are normally optimized with regard to minimum number of steps in the rule, not minimum number of variables. The expert rules work best with closed sets of objects where all possibilities can be included in the training sets. The expert rules can be taken to be specified paths along the perimeter of the multidimensional measurement space (hypercube) to a vertex nearest the SIMCA cylinders for an appropriate class. Overall the performance of the expert system was very good.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL INFORMATION SERVICE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS (This Report) UNCLASSIFIED	21. NO. OF PAGES 20
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

EPA/600/J-88/254
JOURNAL ARTICLE

Analytica Chimica Acta, 211 (1988) 11-29
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CLASSIFICATION OF BINARY MASS SPECTRA OF TOXIC COMPOUNDS WITH AN INDUCTIVE EXPERT SYSTEM AND COMPARISON WITH SIMCA CLASS MODELING

DONALD R. SCOTT

U.S. Environmental Protection Agency, Environmental Monitoring Systems Laboratory, Research Triangle Park, NC 27711 (U.S.A.)

(Received 25th January 1988)

SUMMARY

The performance of an inexpensive, inductive rule-building expert shell system, based on the ID3 algorithm, was compared to that of SIMCA class modeling in classifying the binary mass spectra of 78 toxic and related compounds. The compressed mass spectra consisted of 17 masses chosen by using information theory. The expert rules verified the six main classes and two subclasses found with SIMCA class modeling. These classes were: all benzenes and all alkanes/alkenes (alka(e)nes); nonhalobenzenes, chlorobenzenes, bromoalka(e)nes, and chloroalka(e)nes; and mono-, dichloroalka(e)nes and polychloroalka(e)nes. Training set classification accuracies obtained with the expert system were 93 to 100% as opposed to 62 to 98% for SIMCA. For 73 compounds, the expert rules gave a classification accuracy of 97 to 100% vs. 79 to 96% for SIMCA. Predictive accuracy for the four main classes was 78%. In general, fewer masses were involved with the rules than with the SIMCA models, and the rules are normally optimized with regard to minimum number of steps in the rule, not minimum number of variables. The expert rules work best with closed sets of objects where all possibilities can be included in the training sets. The expert rules represent planes partitioning the multidimensional measurement space (hypercube) into a subvolume nearest the SIMCA cylinders for an appropriate class. Overall, the performance of the expert system was very good.

The present method of choice for identification and quantitation of volatile organic compounds in environmental samples is gas chromatography/mass spectrometry (GC/MS). Because of the voluminous amount of analytical data produced by this technique, chemometric methods have been investigated for efficient extraction of more useful information from GC/MS data files [1–3]. There is a particular interest in developing methods which can be easily implemented by chemists on personal computers. Classification of the low-resolution mass spectra of 78 toxic and related compounds was recently accomplished [1] by using information theory to select key masses from the full binary mass spectra and SIMCA pattern recognition to classify the resulting compressed mass spectra. Classification of the full-intensity mass spectra

and identification of these compounds was also recently accomplished with the use of the autocorrelation transform of the original mass spectra [2,3]. Relatively inexpensive commercial versions of inductive rule-building expert shell systems are now available. Because these software packages are extremely easy to use and the derived decision rules alone often provide valuable information concerning relevant classification variables (features), a study of the applicability of an inductive expert system for the classification of the binary mass spectra of these 78 toxic organic compounds was performed.

Ying et al. [4] have reported the development and testing of an expert system for the interpretation of the infrared spectra of organic compounds in hazardous waste. Derde et al. [5] recently compared the performance of two rule-building expert systems, TIMM and EX-TRAN, with two pattern recognition techniques, *k*-nearest neighbor and linear discriminant analysis, for the classification of olive oils based on gas chromatographic data. They tested for discrimination of two classes with two different data sets. In the present study, an inexpensive and user-friendly expert system, 1ST-CLASS, similar to EX-TRAN was chosen for comparison with the previous SIMCA mass spectral classification results. 1ST-CLASS, like EX-TRAN, uses the ID3 algorithm of Quinlan [6] which produces an optimized decision tree eliminating irrelevant classification variables. This sequential classification method was used to investigate the compressed binary mass spectra of the 78 toxic compounds. This data set contains some eight classes and subclasses and represents a typical classification problem in analytical chemistry with some overlapping of the various classes and some irrelevant variables. The classes resulting from the expert system and the corresponding derived rules are compared with those previously found with the SIMCA class modeling method. It will be shown that the classes found with the expert system are identical to those found with SIMCA modeling and that classification accuracy with the expert system can be as high as 100%.

BACKGROUND

Pattern recognition and measurement spaces

A spectral data vector consisting of p variables, in this case 17 for the compressed mass spectrum, for an object (compound) can be considered to represent a point in a corresponding p-dimensional measurement space. Each variable value (mass channel) represents a coordinate along the appropriate variable axis. With nonbinary data, the limits of this space are represented by a p-dimensional hypercube with an edge length equal to the intensity of the most intense peak in the mass spectrum, the base peak. This length can be set equal to one. The interior, faces, and edges of this hypercube measurement space will be occupied by the object points in the data set as their coordinates

dictate. Distances between points can be calculated from the Euclidean or other metrics.

In the case of binary encoded data, the spectral data vector represents a point in a p-dimensional Hamming measurement space which again can be visualized as a p-dimensional hypercube with unit edge. In this case, the unit length is imposed by the binary encoding. However, because only values of 0 or 1 are assigned to spectral variables, all of the object points lie only on the corners of the hypercube. The interior of the measurement space is not occupied. The distance between two points in the Hamming space is the minimum distance along the edges of the hypercube between the two vertices where the points lie.

A basic assumption in pattern recognition is that similar objects will lie close together and form clusters in the corresponding measurement space and that dissimilar objects will be more widely separated from each other. Frequently, the original measurement variables are not the proper ones to use for the classification problem. Subsets and/or linear combinations of the original variables are used which result in a different space called a feature or pattern space. With binary data the pattern spaces will be Hamming subspaces of the original measurement spaces.

Expert systems [7]

An expert system is a computer system which emulates the knowledge and cognitive processes of a human expert in a particular field, specifically in problem-solving. The rule-building software systems usually consist of a knowledge base and an inference engine with additional modules to handle data input and files, produce reports, and write finished "advisor" systems for a particular application. The knowledge base consists of examples of "results" (classes) occurring with varying sets of "factors" (variables or features) and corresponding rules. The inference engine induces (derives) the rules using some type of efficient algorithm, e.g., ID3 [6]. This type of expert system is called an example-based system. Some expert systems also can generate the knowledge base directly from the rule (a decision tree) alone and are called rule-based systems. Good expert systems also present the reasoning behind the prediction of a given result. Variables which are not important to the required results are eliminated. Individual knowledge bases can be chained together to perform very complex tasks. The "advisor" programs, which are automatically produced by the present software, are series of questions in simple English requesting, in the most efficient order, the values of the variables of importance to the chosen task. Using the derived rule(s), the advisor then produces the results corresponding to the variable input.

In chemical classification problems, the expert system "examples" correspond to the training set data used in supervised pattern recognition. The "factors" correspond to the variables or features of pattern recognition, and the "results" are the classes. The derived rule is of great interest because it con-

tains the variables of importance to the classification problem and the order in which they are applied. The classification approach used is a sequential univariate method rather than a multivariate one.

SIMCA class modeling

The SIMCA (Soft Independent Modeling of Class Analogy) pattern recognition techniques were developed by Wold et al. [8] for classification of chemical data. These class-modeling methods are based on disjoint principal component models for classification of objects. After exploratory data analysis with a principal component projection of the entire data set, classes are determined and training set data chosen for each class. Cross-validation then is used to determine the number of statistically significant principal components for each class. The modeling power of a variable, i.e., the reduction of variance for that particular variable, can be used as a measure of its relevance to a particular class model. Once the class models have been determined, objects are classified by fitting their data to the various class models. A standard deviation for each model is calculated from the training set residuals. This represents a class tolerance distance around the principal component model in measurement space. The standard deviations for the objects from each model are calculated from the corresponding residuals, and the objects are classified based on their distances from the class models. Objects may be found to belong to one particular class or to none of the defined classes. Further details of the technique may be found elsewhere [1–3,8,9]. This method of classification is a multivariate technique and is based on clustering of similar objects in localized regions of the measurement space.

EXPERIMENTAL

Data set

A list of the 78 individual compounds is given in Table 1. The low-resolution mass spectra of the compounds were obtained from the EPA-NIH Mass Spectral Library on an INCOS data system. A typical spectrum contained approximately 16 peaks due to reduction with an INCOS algorithm. The range of mass/charge ratios was from 35 to 256 with 151 different peaks occurring in the total set. The data were binary-encoded with a threshold level of 1% of the base peak intensity. The information content of the 151 different nonzero intensity mass channels was calculated from the binary-encoded mass spectra of the set of 78 compounds. Only those 17 mass channels with the highest information contents were used in producing the compressed mass spectra for the classification study [1]. These key masses are listed in Table 2.

TABLE 1

Compounds included in this study

1. *p*-Xylene
2. 1,3,5-Trimethylbenzene
3. Isopropylbenzene
4. n-Butylbenzene
5. 1-Methyl-4-isopropylbenzene
6. *o*-Dichlorobenzene
7. *p*-Dichlorobenzene
8. 1-Chloro-2-methylbenzene
9. 1-Chloro-4-methylbenzene
10. *p*-Chlorostyrene
11. 1,1-Dichloroethane
12. 1,1,1,2-Tetrachloroethane
13. 1,2,3-Trichloropropane
14. 3-Chloropropene
15. 2-Chlorobutane
16. 1,3-Dichlorobutane
17. 1,4-Dichlorobutane
18. 1,4-Dichloro-2-butene (*cis*)
19. 3,4-Dichlorobutene
20. 1,4-Dioxane
21. 1-Chloro-2,3-epoxypropane
22. 2-Chloroethoxyethene
23. Acetophenone
24. Benzonitrile
25. Benzene
26. Toluene
27. *o*-Xylene
28. *m*-Xylene
29. Ethylbenzene
30. Styrene
31. Chlorobenzene
32. Bromobenzene
33. *m*-Dichlorobenzene
34. 1-Chloro-3-methylbenzene
35. Chloroform
36. Carbon tetrachloride
37. Bromochloromethane
38. Bromotrichloromethane
39. Dibromomethane
40. Bromoform
41. 1,2-Dichloroethane
42. 1,1,1-Trichloroethane
43. 1,1,2-Trichloroethane
44. 1,1,2,2-Tetrachloroethane
45. Pentachloroethane
46. 1,1-Dichloroethene
47. Trichloroethene
48. Tetrachloroethene
49. Bromoethane
50. 1,2-Dibromomethane
51. 1-Chloropropane
52. 2-Chloropropane
53. 1,2-Dichloropropane
54. 1,3-Dichloropropane
55. 1-Bromo-3-chloropropane
56. 1,2-Dibromopropane
57. 2,3-Dichlorobutane
58. Tetrahydrofuran
59. Benzaldehyde
60. 1-Bromo-1-chloroethane
61. 2,2-Dibromopropane
62. 2-Bromopropene
63. 2-Bromopropane
64. 3-Bromopropene
65. 1-Bromopropane
66. 1-Chlorobutane
67. 1-Bromo-2-chloroethane
68. Bromodichloromethane
69. 1-Bromobutane
70. 2,2-Dichlorobutane
71. Dibromochloromethane
72. 1,1,2-Trichloropropane
73. 1,3-Dibromopropane
74. 1,1,1,2-Tetrachloropropane
75. 1,2,2,3-Tetrachloropropane
76. 1,3-Dibromobutane
77. 1,1,2,3-Tetrachloropropane
78. 1,4-Dibromobutane

TABLE 2

Seventeen key masses with highest information content for 78 compounds

Mass	Binary information content (bits)	Mass	Binary information content (bits)
41	0.92	77	0.94
49	0.98	78	0.84
50	0.84	79	0.96
51	0.99	81	0.80
61	0.82	91	0.92
62	0.95	93	0.87
63	1.00	95	0.90
65	0.92	107	0.82
75	0.89		

Hardware and SIMCA software

The SIMCA software package is modular, menu-driven and graphically orientated. Details of the software are given elsewhere [1,8]. The previous SIMCA classification [1] was performed on an Osborne 1 microcomputer (Z80A) with a CP/M operating system and 64K memory. The SIMCA model derivation and classification of the multichloroalka(e)nes (see below) were performed on an IBM AT computer with 512K memory. The expert system was also used on the IBM AT.

Expert system

The expert system used was 1ST-CLASS, version 3.5. This program [10] is very user-friendly with a menu format and many help screens. It is composed of four general modules on different screens: file handling ("Files"), result and example (data) entry ("Definitions"), rule-building methods ("Methods"), and the automatically written expert advisor ("Advisor"). External data files can be imported and reports can be automatically generated. The rule-building algorithm (ID3) is very fast and various modifications to a resulting rule can be made or different rules can be produced from a given set of examples in a "what if" fashion. The classification algorithm produces a decision tree (the rule) optimized for sequential ordering of only relevant variables. The tree is displayed in detail graphically and can be tested and easily debugged. Different knowledge bases can be forward and backward chained (connected) to create very complex expert systems.

Data entry. Manual entry of results and examples is easily accomplished with a spreadsheet type of screen with editing capabilities. Columns are factor values (text or numeric data) and rows are different examples. The last column is the set of allowed results (text). A maximum of 32 factors, 32 results

and 255 examples per knowledge base is allowed by the program. In addition to the actual data entered, a weight must be assigned for each example. The default is one, but values of 0 to 100 are allowed. A weight of zero does not remove an example from consideration during rule derivation. This is accomplished with a deactivation designation which temporarily removes designated examples and/or factors from consideration. The weights are used to compute "statistics" which are displayed with the derived rule and are useful for testing and analyzing the rule.

Rule building. There are four possible options for creating rules with this system: optimize, left-right, customize, or match. The last two are not inductive methods. The customize method is used to input an already known rule and to test and edit it. The match method is used with sets of examples with very little underlying structure which generate large unwieldy rules. In this case, matching of factors against examples in the knowledge base may provide useful advice. The first two methods are inductive ones and are of primary interest for the present classification study. The optimize method produces a decison-tree rule from the given examples, omitting redundant and irrelevant factors. The result is a graphic display of the ordered factors leading to the various types of results. The factors are displayed in the most efficient order to solve the classification problem as determined by the alogrithm, ID3. The left-right method is a variant of the optimize method where the order of the factors in the decision tree is determined, not by classification importance, but by their position in the data table from left to right. In addition to the decision tree there is displayed on the rule screen a table of statistics for the rule. This includes numbers of total active examples; number and designation of active examples leading to a given result; frequency of occurrence of a given result; proportion of total active examples giving a particular result; and the probability of getting a particular result when there is more than one result at a node. This information is useful when testing the result.

ID3 algorithm. The algorithm which is used to produce the optimized decision tree is the Iterative Dichotmiser, version 3 (ID3) of Quinlan [6]. ID3 selects from all of the factors in the examples the one which is the most discriminating for the set of active examples. It then picks the next most discriminating, continuing this process until all of the factors of importance to the classification are selected. Factors which are not necessary to accomplish the classification are ignored. The method used to select the ordering and number of factors is based on information-theory concepts [6]. The decision tree is regarded as an information source which produces a message, the class of a given object. The complexity of the decision tree is assumed to be strongly related to the amount of information conveyed by the message. The construction of the decision tree is an iterative process initiated with a subset of examples. A rule is induced for this subset, using the information-theory concepts to select factors. A new rule is derived using the original subset and exceptions

to the first rule. This process is repeated until no exceptions to the latest rule are found. The result is a compact decision tree containing only relevant and nonredundant variables. ID3 was designed to handle very large numbers of examples and its computation time increases only linearly with the product of the number of examples, number of factors, and number of nodes. It has been tested for a set of about 2000 objects and 14 factors resulting in a decision tree with 48 nodes.

Advisor. The advisor program presents a series of questions to the user in simple English regarding the values of the factors found in the rule-building phase to be important in producing the specified results. These questions are asked in the order specified in the decision tree. After responses are input for the appropriate number of queries, the advisor automatically produces a result, which is obtained from the derived rule.

Data analysis

The binary-encoded compressed mass spectra of the 78 compounds composed of only the 17 mass channels found to have the highest information content for the set of compounds were used in both the SIMCA and expert-system classification. After exploratory analysis of the entire data set by principal component projections, classes were defined and training sets chosen. All objects in a given SIMCA class were used for the training sets, except for bromobenzene which was only used in the total aromatic training set. Five compounds including bromobenzene were not included in the training sets but were used as test objects. These were *p*-dioxane, tetrahydrofuran, 1-chloro-2,3-epoxypropane, 2-chloroethoxyethene, and bromobenzene. The number of compounds used in the training sets for the different SIMCA classes is shown in Table 3. Cross-validation was used to determine the number of principal components in each SIMCA class model. Modeling power was used to eliminate

TABLE 3

Number of compounds in chemical classes and training sets[a]

Substituent/Class	Aromatics	Alka(e)nes[b]	Others[c]
None	14 (14)	0	2
Br	1	14 (21)[d]	0
Cl	8 (8)	30 (30)	2
Br and Cl[e]	0	7[d]	0
Total	23 (23)	51 (51)	4

[a]Numbers in parentheses are numbers of compounds used in training sets. [b]Alka(e)nes includes alkanes and alkenes. [c]"Others" includes ethers and epoxides. [d]The seven bromo- and chloroalka(e)nes were used in the training set for the bromoalka(e)ne class. [e]Substituted with both chlorine and bromine.

irrelevant variables in these models. An individual mass spectrum was classified as belonging to the SIMCA class to which the spectrum was closest. Further details are given elsewhere [1].

The multichloroalka(e)ne (tri-, tetra-, or pentachloroalkanes/alkenes) SIMCA class was determined in a similar manner with a modeling power limit of 0.18 or greater for variable inclusion. The training set included all of the fourteen appropriate alkanes/alkenes. After the model was established, all 78 compounds were classified relative to the model.

All of the compressed binary mass spectra of the 78 compounds were entered into the data table of the expert system and subsets were classified by inactivating appropriate examples. The compressed spectra resulted in 78 total examples for the expert system with the 17 mass channels as factors for each example. The same 73 compounds or subsets were used in the training sets as were used for the SIMCA models. All 17 masses were available for each class rule determination. Weights for active examples were the default values of one. The results input to the system consisted of the suspected classes for each set of active examples. The rule-building mode normally used was the optimize method described above. Classification of the test spectra was accomplished by using the resulting advisor modules for each derived rule. Only in the derivation of minimum variable rules were any of the 17 mass variables withheld from use in the rule. Generally, only rules which achieved complete separation of all training spectra were considered successful. In the classification of the mono- and dichloroalka(e)nes vs. the tri-, tetra-, and pentachloroalka(e)nes, the remainder of the 73 spectra were classified together as a third "others" class and 93 and 94% separation of the training data was accepted.

RESULT AND DISCUSSION

SIMCA models

In the previous SIMCA study [1] of the binary mass spectra of these compounds, six main classes and one subclass of the chloroalka(e)nes were found. Alka(e)nes is used here to designate both alkanes and alkenes because these types of compounds can not be distinguished from each other in this study. The compounds of interest are listed in Table 1. Seventy-three of these were used in the training sets for the various classes. Five of the target compound spectra were used as test spectra: *p*-dioxane, tetrahydrofuran, bromobenzene, 1-chloro-2,3-epoxypropane, and 2-chloroethoxyethene. The six main classes consisted of all alka(e)nes, all benzenes, nonhalobenzenes, chlorobenzenes, chloroalka(e)nes and bromoalka(e)nes. The bromoalka(e)nes class contained both bromoalka(e)nes and bromochloroalka(e)nes. The single subclass consisted of the mono- and dichloroalka(e)nes. The 17 key masses with highest information content for the set of compounds were used in the classification. These are listed in Table 2. The numbers of compounds in each class

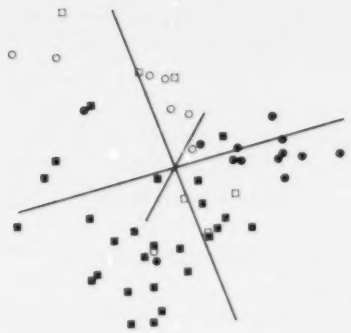

Fig. 1. Three-dimensional principal component plot of mass spectra of toxic and related compounds: (□) chlorobenzenes; (○) nonhalobenzenes; (■) chloroalka(e)nes; (●) bromoalka(e)nes.

TABLE 4

SIMCA class model for polychloroalka(e)nes[a]

	Principal component 1					
Training set	14 compounds					
Mass channel	41	51	63	75	77	78
Loadings	0.45	0.38	0.32	0.42	0.42	0.45
Model power	0.49	0.29	0.19	0.39	0.39	0.48
Remaining variance	41%					
Residual s.d.	0.70					
Classification accuracy						
Training set	71%					
All compounds	80%					

[a]This includes all trichloro-, tetrachloro-, and penta-chloroalka(e)nes. No bromochloroalka(e)nes are included.

training set are given in Table 3. The SIMCA class models consisted of only one principal component with the exception of the all alka(e)nes model which had two. The all alka(e)nes model had twelve masses, and all others had five. SIMCA classification accuracy of the spectra was 96% for the all alka(e)nes vs. all benzene classes; 82% for the nonhalobenzenes, chlorobenzenes, chloroalka(e)nes, bromoalka(e)nes classes; and 91% for the mono- and dichloroalka(e)nes subclass. A three-dimensional principal component plot of the 78 compound spectra coded for the four main classes is shown in Fig. 1.

Further investigation of the fourteen tri-, tetra-, and pentachloroalka(e)nes with SIMCA in this study yielded a single principal component model for these spectra with a remaining variance of 41% and a residual standard deviation of

0.70. Six masses were found to be important in this model: 41,51,63,75,77, and 78. The classification accuracy using a residual standard deviation of 0.76 was 71% for the training set and 80% for all compounds. A summary of the model is given in Table 4.

Expert rule models

There is no simple way to perform exploratory data analysis with the expert system other than a trial-and-error method of guessing classes and trying to construct successful rules for these classes. Therefore the eight classes found in the SIMCA study were used as the result for the expert system. This also allowed a fair method of comparing the two methods in a supervised pattern recognition mode of operation. The expert system uses a partitioning method and the examples should be an exhaustive closed set for best classification results. However, for comparison of the SIMCA and expert rule methods, exactly the same numbers of classes and training set compounds were used for both methods. The exception was the two chloroalka(e)nes subclasses where another class labeled "others" consisting of 43 compounds in addition to the 30 chloroalka(e)nes was used. The five special test spectra were always withheld from the training sets except for bromobenzene which was used in the all-benzenes model. These test spectra were classified by using the advisor for the appropriate rule.

TABLE 5

Expert system derived rules for various classes[a]

Class	Compounds in training set	No. of masses	Lines in rule[b]	No. of singlets[c]
All alka(e)nes[d]	51	6	17	1
All benzenes	23	6	17	1
Nonhalobenzenes	14	10	29	0
Chlorobenzenes	8	10	29	1
Chloroalka(e)nes	30	10	29	3
Bromoalka(e)nes	21	10	29	1
Mono- and di-chloroalka(e)nes	16	13	40	2
Tri-, tetra-, and pentachloroalka(e)nes	14	13	40	5
Others[e]	43	13	40	4

[a]These rules are optimized for fewest number of lines. [b]This is an indication of the complexity of the rule. [c]Number of lines in rule which result in classification of only one example. This is an indication of individual exceptions to the more general class rules. [d]Alka(e)nes include both alkanes and alkenes. [e]Others include all other compounds which were not chloroalka(e)nes, except for the five test compounds.

TABLE 6

Relevant masses for SIMCA models and expert system classes

Class	SIMCA masses	Expert system masses[a]
All alka(e)nes	41,49,51,61,62,63, 75,77,79,81,93,107	49,50,62,65,77,91
All benzenes	50,63,65,75,91	49,50,62,65,77,91
Nonhalobenzenes	49,50,65,75,91	49,50,62,63,75,77,81,91,93,107
Chlorobenzenes	50,62,63,65,91	49,50,62,63,75,77,81,91,93,107
Chloroalka(e)nes	41,51,65,77,95	49,50,62,63,75,77,81,91,93,107
Bromoalka(e)nes	49,61,62,63,65	49,50,62,63,75,77,81,91,93,107
Mono- and di-chloroalka(e)nes[b]	63,65,75,77,91	41,49,61,62,63,65,77,78,79,91, 93,95,107
Tri-, tetra-, and pentachloroalka(e)nes	41,51,63,75,77,78	41,49,61,62,63,65,77,78,79,91, 93,95,107

[a]These masses appear in the rule for differentiating between the classes. The rules are optimized for fewest number of nodes. [b]The classification of these chloroalka(e)nes was treated as a three-class case: mono- and di-chloroalka(e)nes; tri-, tetra-, and penta-chloroalka(e)nes; and others.

The derived rules and results for the eight classes are summarized in Tables 5, 6 and 7. Successful expert rules were derived for all of the eight SIMCA classes with 100% classification accuracy for all training set spectra except the chloroalka(e)ne subclasses. The SIMCA training set accuracies ranged from 79 to 98% for the six main classes. The expert rule accuracies for the chloroalka(e)ne subclasses were 93 and 94% for training spectra. For these subclasses, the SIMCA training set classification accuracies were only 62 and 71%. The resulting expert rules contained only six, ten, and thirteen of the 17 key masses and consisted of 17, 29, and 40 lines per rule. The chloroalka(e)ne subgroups gave the most complicated rules. The number of branches in the decision tree (rule) which resulted in only one compound spectrum being classified (singlets) ranged from zero for nonhalobenzenes to five for tri-, tetra-, and pentachloroalka(e)nes.

The rule (decision tree) for the four main classes is shown in Fig. 2. The numbers in parentheses at the right of the figure are the number of training set compounds classified with that particular branch of the tree. The major branch for the nonhalobenzenes (benzenes), which included nine of the fourteen training compounds, involves the sequential absence of masses 93 and 49, the presence of masses 77 and 63, and the absence of mass 62. The major branch for the chlorobenzenes, which includes five of eight training compounds, involves the sequential absence of masses 93, 49, 77, and the presence of mass 75. For the chloroalka(e)nes, the major branch, which classified eighteen of thirty training compounds, involved the sequential absence of mass 93, the

TABLE 7

Comparison of classification accuracy of target compounds by SIMCA modeling and expert rule-building system[a]

Classes	Classification accuracy (%)			
	Training set[b]		All compounds[c]	
	SIMCA	Expert	SIMCA	Expert
All benzenes	91	100	96	100
All alka(e)nes[d]	98	100	96	100
Nonhalobenzenes	79	100	93	100
Chlorobenzenes	87	100	94	100
Chloroalka(e)nes	86	100	84	100
Bromoalka(e)nes[e]	80	100	93	100
Mono- and di-chloroalka(e)nes	62	94	92	97
Tri-, tetra-, and penta-chloroalka(e)nes	71	93	79	97

[a]The SIMCA results are taken from reference 1 except for the multichloroalka(e)ne class. The rules are optimized for fewest nodes. [b]The numbers of compounds in each training set are listed in Table 3. [c]This includes all compounds except the special test compounds. [d]Alka(e)nes is used to designate both alkanes and alkenes. [e]Bromoalka(e)nes includes both bromoalkanes/alkenes and bromochloroalkanes/alkenes.

presence of mass 49, and the absence of mass 50. For the bromoalka(e)nes, the major branch, which classified seventeen of twenty-one training compounds, involved the sequential presence of masses 93 and 81.

The five test-compound spectra provide a particularly difficult test for the expert rules because these were not included in their derivation and they include two cyclic alkylethers, a halobenzene containing bromine, a chloroalkyl epoxide, and a chloroalkenyl ether. These compounds do not exactly fit into any of the derived classes. The classification results from the expert rules as well as from SIMCA modeling are listed in Table 8. Both SIMCA and the expert rules place the four compounds containing alka(e)ne structures correctly in the all-alka(e)nes class as opposed to the all-benzene class. In the category of the four main classes, both methods place the compounds into correct skeletal groups, i.e., bromobenzene into a benzene class and the others into an alka(e)ne class. Furthermore, both methods correctly identify the chlorine-containing alka(e)nyl epoxide and ether as chloroalka(e)nes. However, SIMCA classifies the two cyclic ethers as bromoalka(e)nes (i.e., alka(e)nes but not chloroalka(e)nes?), while the expert rules classify them as chloroalka(e)nes. Bromobenzene is classified as a chlorobenzene (i.e., a ben-

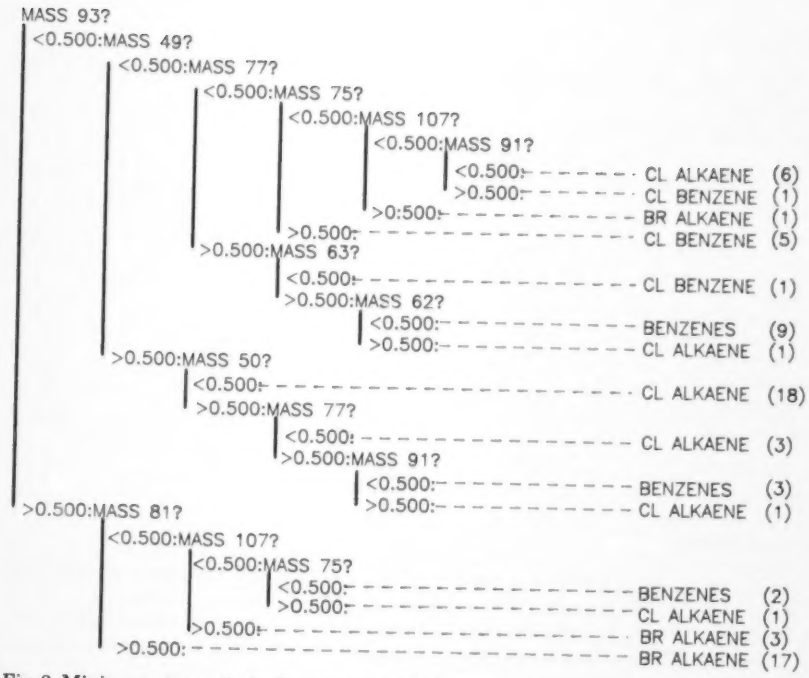

Fig. 2. Minimum step optimized expert rule for four main classes.

zene but not a nonhalobenzene?) by SIMCA, and the expert rule assigns it to the nonhalobenzene class. In the case of the determination of the number of chloro substituents on the test compounds, the expert system correctly places the two monochloroalka(e)nyl compounds into the correct class and places the nonchloro compounds in the "others" class. SIMCA classifies the three nonchloro compounds as belonging to neither of the two defined chloroalka(e)ne classes. SIMCA correctly classifies the monochloroalkyl ether as a mono- or dichloroalka(e)ne, but incorrectly classifies the monochloroalkyl epoxide as not belonging to either of the chloroalka(e)ne subclasses. This latter spectrum was closer to the mono- or dichloroalka(e)ne subclass than to the polychloroalka(e)ne subclass but was at a distance of 0.98 from the model which is larger than the two subclass limits of 0.73 or 0.76.

Comparison of SIMCA and expert system

The expert system rule method is a sequential univariate partitioning (discriminating) method, while the SIMCA technique is a class modeling and discriminating method. The expert rules work by sequentially partitioning the set of all objects into appropriate subsets of classes until all objects have been

TABLE 8

Classification of special test compounds by expert rules and SIMCA[a]

Class	Bromobenzene	*p*-Dioxane	Tetrahydro-furan	1-Chloro-2,3-epoxypropane	2-Chloroeth-oxyethene
All alka(e)nes		X(S)	X(S)	X(S)	X(S)
All benzenes	[b]				
Nonhalobenzenes	X				
Chlorobenzenes	(S)				
Chloroalka(e)nes		X	X	X(S)	X(S)
Bromoalka(e)nes		(S)	(S)		
Mono-, di-chloroalka(e)nes				X	X(S)
Tri-, tetra-, penta-chloroalka(e)nes					
Others	X(S)	X(S)	X(S)	(S)	

[a]X designates the expert system result. S denotes the SIMCA result. [b]Bromobenzene was used in the all benzenes training set.

uniquely classified, if possible. If the inherent structure is not present in the set of objects with the given variables, then the rule will show multiple classes for some objects. SIMCA classifications are based on constructing class models in the multivariate pattern space with spatial limits and comparing the distances of a given object to the various class models. If the distance of an object is within the limits of a class model, then the object is classified as belonging to that class. An object can belong to none of the classes by exceeding the class model limits for all defined models or can belong to more than one class by lying within more than one class limit. SIMCA discriminates by partitioning the set of objects into subsets of objects fitting each modeled class and another set of objects which do not fit any defined class (i.e., "others"). Therefore the objects belonging to none of the defined classes are automatically identified with SIMCA.

In the case of the expert rules, examples of all classes, including the "others" category, theoretically should be included during derivation of the rule. This is a weakness of the expert rules because it is clearly impossible to enumerate every example which might be encountered. However, as seen from the present test spectra results, useful results can be obtained without listing every example. The best procedure to circumvent this weakness is to include common counter examples in the rule-building phase. For example, in the present case the examples with all seventeen masses present (all one's) and all seventeen masses absent (all zero's) should be included. These do not correspond to any real compound spectra in the target set and could be assigned a class of "unknown" or "others".

Minimum variable expert rules

The total number of masses actually used in the derived rules for the various classes are of interest. They are collected in Table 6 along with those for the corresponding SIMCA models. The expert rules contain only one set of variables for each group of classes, while SIMCA requires one set for each class model. Generally, one would expect that fewer variables would be required to partition classes than to model and partition classes as with the SIMCA technique. This is supported by the smaller number of masses for the expert rules than for the sum of the corresponding SIMCA classes. The exception is the case of the chloroalka(e)ne subclasses where thirteen masses were used for the expert rule and only eight for the two SIMCA classes. This may be due to the expert rule partitioning three classes (the two chloroalka(e)ne subclasses and others), while there were only two explicit SIMCA class models. All of the seventeen masses were used in at least one of the eight SIMCA models. The expert rules used all of the seventeen masses except mass 51.

The expert rules derived by the ID3 algorithm in this software implementation are not optimized for the minimum number of variables to partition a set of objects. They are optimized to partition with a minimum number of steps (lines) in the decision tree which may not result in the minimum number of variables. The derived rules also are not unique if the requirement of minimum steps is relaxed. The results for the four main class problem in Table 9 illustrate this point.

TABLE 9

Expert rules for four major classes by using minimum number of masses[a]

	Minimum variable rules[b]				Minimum node rule[c]
	1	2	3	4	
Number of masses	8	8	8	8	10
Number of lines	31	33	33	35	29
Pertinent masses[d]		(93, 91, 77, 63, 50, 49) and			49,50,62,63,75, 77,81,91,93,107
	62,107	75,107	62,81	75,81	
Classification accuracy					
Training set (%)	100	100	100	100	100
All compounds (%)	100	100	100	100	100

[a]Benzenes, chlorobenzenes, chloroalka(e)nes, and bromoalka(e)nes. [b]The minimum variable rules were determined to achieve complete classification of the training set and to have the lowest number of masses possible. [c]This rule is the one obtained directly from the algorithm with 17 available masses. [d]All minimum variable rules have the six masses in parentheses in common and the other two additional ones.

The rules presented previously for the four major classes are those produced with the optimize mode of the software with all seventeen masses available for use. Alternative rules derived for minimum numbers of variables were sought for this set of classes. The software-optimized rule had ten masses and 29 lines. One variable at a time was deleted from these particular ten masses and new rules derived for the four classes. It was found that seven other rules with only nine variables and numbers of lines ranging from 29 to 33 could be found to separate completely all examples into the proper classes. It was also found by single mass deletion from the original optimized rule that masses 93, 77, and 63 were critical to complete separation. With these three masses as a base, others from the original tree were added until a combination was found which would separate all the examples correctly. Four different rules with only eight masses and 31 to 35 lines were found to completely separate the examples. These rules are summarized in Table 9.

All of these four rules, which only use eight masses, consist of more lines than the minimum node rule produced in the software optimize mode. All four have six masses in common (93, 91, 77, 63, 50, 49) and a pair of masses including either 62 or 75 and 81 or 107. The rule with the fewest lines (31) is the one using the basic six and 62 and 107. These four possible rules have the same classification accuracy as the minimum node rule. This shows that the rules derived with the optimize mode are not necessarily unique with regard to successful classification and are definitely not minimum variable rules.

Predictive classification accuracy

Because the expert rules performed very well in classifying the examples in the training and test sets, the rule for the four main classes was further tested for predictive accuracy. Ten percent or a minimum of two compounds from each class were randomly selected and withheld from the training sets along with the five special test compounds. An expert rule with all seventeen masses available was then derived by using the optimize mode. This particular rule involved nine masses (mass 49 was not included) and 27 lines as compared to the original rule involving ten masses and having 29 lines. All of the withheld spectra were then classified by using this rule. The classification accuracy of the withheld spectra, excluding the five special test spectra, was 78%. The two chloro test compounds were classified as chloroalka(e)nes as with the original rule. However, bromobenzene was classified as a bromoalka(e)ne as opposed to a nonhalobenzene with the original rule. Also, both cyclic ethers were incorrectly classified as chlorobenzenes as opposed to chloroalka(e)nes with the original rule. Similar classification accuracies were obtained by withholding 20 or 33% of the training set spectra. The difference between the essentially 100% accuracy with all the training set spectra included and the lower ca. 80% obtained here is probably due to the heterogeneous nature of the individual spectra within the classes.

Conclusions

The expert rules derived for the set of binary-encoded mass spectra of 78 toxic and related compounds verified the existence of the six main classes and two subclasses in the data set found with SIMCA class modeling. The training set classification accuracies obtained with the expert system were 93–100% as opposed to 62–98% for SIMCA. For the 73 compounds which definitely fit into one of the classes the expert rules gave a classification accuracy of 97–100% vs. 79–96% for SIMCA. The lowest accuracies were obtained with the two chloroalka(e)ne subclasses. Predictive accuracy for the four main classes obtained by withholding some of the training set spectra was a lower 78%. Rules in general involved fewer masses than did the SIMCA models. The rules obtained with the software optimize mode are optimized with regard to minimum number of steps in the rule, not minimum number of variables. Alternative rules to the one produced in the optimize mode can be produced as shown above and sometimes these other rules involve fewer variables than the original one.

SIMCA models for these mass spectra can be visualized as cylinders for the single principal-component models localized around the centroids of the various classes in the 17 dimensional Hamming pattern space (hypercube). If the class cylinders do not overlap, then separation of the classes is ensured and modeling will also produce partitioning. The expert rules can be interpreted as specific planes partitioning the hypercube into a subvolume nearest the SIMCA cylinders for an appropriate class.

The expert shell system tested, 1ST-CLASS, was inexpensive, fast, and very easy to use. It has limits of 32 for numbers of variables and classes, but chaining can frequently solve that problem. The rules derived using the ID3 algorithm will work best with closed sets of objects to classify where all possibilities can be included in the examples (training sets). However, this is also true for SIMCA or other pattern recognition techniques. One advantage of SIMCA modeling is that exclusion of an object from all defined class models is automatically considered. This is not true with the partitioning expert rules, and for best results this "others" class must be included. Overall, the performance of the expert system on this set of binary mass spectra vs. SIMCA was very good. These convenient and inexpensive expert systems may well satisfy the needs of chemists for many of their analytical classification problems.

Although the research described in this article was funded by the United States Environmental Protection Agency, it has not been subjected to Agency review. It therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

REFERENCES

1 D. R. Scott, Anal. Chem., 58 (1986) 881.
2 D. R. Scott, W. J. Dunn III and S. L. Emery, Environ. Sci. Technol., 21 (1987) 891.
3 W. J. Dunn III, M. G. Koehler, S. L. Emery and D. R. Scott, Chem. Lab., 1 (1987) 321.
4 L. S. Ying, S. P. Levine, S. A. Tomellini and S. R. Lowry, Anal. Chem., 59 (1987) 2197.
5 M. P. Derde, L. Buydens, C. Guns, D. L. Massart and P. K. Hopke, Anal. Chem., 59 (1987) 1868.
6 J. R. Quinlan, in R. S. Michalski, J. F. Carbonell and T. M. Mitchell (Eds.), Machine Learning: An Artificial Intelligence Approach, Tioga Publishing, Palo Alto, CA, 1983, pp. 461–482.
7 P. Jackson, Introduction to Expert Systems, Addison-Wesley, Reading, MA, 1986.
8 See, e.g., S. Wold, C. Albano, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjostrom, in B. R. Kowalski (Ed.), Chemometrics: Mathematics and Statistics in Chemistry, D. Reidel, Boston, 1984, pp. 17–96.
9 M. A. Sharaf, D. L. Illman and B. R. Kowalski, Chemometrics, Wiley, New York, 1986, pp. 242–257.
10 W. Thomas and W. Hapgood, 1ST-CLASS Instruction Manual, Programs in Motion, Wayland, MA, 1987.

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.